UNIVERSITÀ DEGLI STUDI DI PADOVA
Dipartimento di Territorio e Sistemi Agro-Forestali (TeSAF)

Corso di laurea magistrale/Second Cycle Degree (MSc)
in Food and Health

Relatore/Supervisor
Prof. Giovanni Franzo

Laureanda/o /Submitted
by Anna Belous

Matricola n./Student n.
2109247

ANNO ACCADEMICO/ACADEMIC YEAR 2024/2025

Abstract

Hepatitis E virus (HEV), particularly genotypes 3 and 4, represents a significant zoonotic threat with complex transmission dynamics involving humans, domestic animals, and wildlife. This thesis explores the application of neural networks to predict the host origin of HEV based on viral genomic sequences, utilizing a bioinformatic approach centered on k-mer encoding of nucleotide data. A curated dataset of 448 complete HEV genomes with verified host metadata was assembled, focusing on zoonotic genotypes 3 and 4 to uncover subtle genomic signatures indicative of host tropism.

A fully connected feed-forward neural network was developed and trained on normalized 6-mer frequency vectors, with performance evaluated through stratified 5-fold cross-validation and independent testing. The model achieved moderate predictive accuracy (~20–70%), excelling in classifying well-represented hosts such as humans and rats, while struggling with underrepresented wildlife hosts due to data imbalance and biological similarity. Phylogenetic analysis corroborated the zoonotic potential of HEV, revealing frequent cross-species transmissions and highlighting critical transmission pathways including foodborne, occupational, and environmental routes.

This study demonstrates the feasibility of integrating machine learning with viral genomics to enhance zoonotic risk assessment. Despite current limitations related to data scarcity and class imbalance, the findings emphasize the promise of computational methods for improving surveillance and control strategies for HEV outbreaks. Future work focusing on data enrichment, hierarchical classification, and incorporation of richer genomic features could further advance predictive accuracy and public health impact.

1. INTRODUCTION

1.1. Overview of hepatitis E virus

Hepatitis E virus (HEV), especially its genotypes 3 and 4, is a significant zoonotic infection posing a public health problem. It causes acute hepatitis in humans and animals, both in the European Union and in developing countries. Its zoonotic potential and ability to be transmitted through food and water make it particularly relevant for study in the context of globalization and active international trade. Evolutionary changes in the virus and its genetic variability can influence the pathogenicity and spread of disease. This underscores the need for in-depth analysis of its epidemiology, tropism determinants, and phylogenetic characteristics for further improving the control of the virus. Virologists, epidemiologists, and veterinarians are the main players in the study of HEV, making a huge contribution to understanding its evolution and spread. They examine the dynamics of the virus's spread through factors such as changes in ecosystems, animal migration, changes in the population's eating habits, hygiene, and the keeping and care of farm animals.

HEV is characterized by remarkable genetic variability, which, as previously mentioned, has led to its classification into different genotypes. Although belonging to the same viral species, these genotypes display markedly different biological and epidemiological behaviors. As documented cases show, genotypes 1 and 2 are most common in developing countries, with most outbreaks occurring in Asia. These genotypes are transmitted via the fecal-oral route, and most of the reported cases are associated with the consumption or use of contaminated water for domestic purposes (Khuroo et al., 2016). Genotypes 3 and 4, on the other hand, are often associated with zoonotic transmission. Cases often describe the consumption of undercooked pork or wild boar meat. The confirmed link between hepatitis E and the consumption of undercooked or insufficiently cooked meat underscores the need to expand methods of informing the public about the risks associated with hepatitis E and methods of its prevention. The impact of HEV varies significantly from region to region, with genotype 3 being particularly problematic in Europe and North America (Songtanin et al., 2023). In these settings, the burden of disease is often underestimated, as many cases go unreported or misdiagnosed. Thus, it is critical to grasp the nuances of demographic and geographic factors when evaluating HEV transmission dynamics. Such understanding is vital for tailoring public health responses and surveillance strategies effectively. Transmission pathways for HEV are diverse, and understanding these pathways is fundamental for mitigating its spread. Animal reservoirs, such as swine or wild boar, play a pivotal role in the zoonotic transmission of HEV to humans, often leading to periodical outbreaks or isolated cases in populations with close contact to infected animals. This host-pathogen interaction is also influenced by ecological factors, which can enhance the likelihood of zoonotic outbreaks. Due to this dynamics it’s crucial now to integrate animal and human health in one approach. A focus on the epidemiological interactions between human and animal populations should provide us insights on potential interventions that could prevent outbreaks. Other genotypes, have been detected in other species: genotype 5 and 6 have been detected in wild boars in Japan, while infections with genotypes 7 and 8 have been identified in Middle Eastern and Chinese populations of dromedaries and Bactrian camels (Wang et al., 2019).

In addition, the replication mechanisms of hepatitis E in both humans and animals are strongly influenced by various interactions between the host and the pathogen. It can cause acute and, in certain cases, chronic inflammation and disease, which is particularly dangerous for people with weak immune systems or who do not have access to timely modern medicine. This feature of hepatitis E is a problem for the healthcare systems of developing countries, since hepatitis E in its chronic stage can lead to irreversible severe liver disease (Mirazo and Arbiza, 2019). This problem has been considered by the WHO for several years, but no vaccine for hepatitis E has yet been developed. This fact particularly highlights the importance of preventive measures and computer-based methods for assessing the risk of hepatitis E. It is also necessary to conduct more research on the various pathways of virus replication, the role of host immunity, and the factors that contribute to the development of chronic infection.

1.1.1. Genomic structure of HEV

The key to understanding the classification, pathogenicity, and mutation potential of hepatitis E is its genomic structure. It is a single-stranded RNA virus with positive polarity, and its genome is typically 7.2 to 7.5 kilobases in length, encoding several proteins that are essential for virus replication and hepatitis E infection pathogenesis . Three open reading frames can be identified in the HEV genome: ORF1, ORF2, and ORF3. ORF1 is responsible for encoding non-structural polypeptides, such as RNA polymerase, which are involved in virus replication. ORF2 encodes the viral capsid protein, which plays a crucial role in the formation of viral particles and interaction with the immune system. ORF3 is associated with the encoding of a multifunctional protein involved in governing the egress process from infected cells. and more generally contributing to the pathogenesis of the hepatitis E virus (Kamani et al., 2021).

At present, at least eight genotypes of the hepatitis E virus are known. Genotype 1 is most commonly found in humans and is associated with large recorded outbreaks of waterborne infection. It is endemic in developing countries, especially in Asia, Africa, and Mexico. Genotype 2, like genotype 1, is also associated with human infections. It is less common and is most often found only in African countries and Mexico. Genotypes 3 and 4 are the focus of this study because of their ability to infect humans through contact with animals. Surprisingly, genotype 3 is most commonly found in developed countries, especially in Europe and North America. According to documented cases, it is associated with the consumption of pork or other meat products that have not undergone sufficient heat treatment. Genotype 4 is also zoonotic, but differs from genotype 3 in its geography. It is widespread in Asia, especially in China, and is also associated with the consumption of pork. For completeness, other genotypes have been identified; i.e. 5, 6, and 7. Genotype 5 is mainly found in wild animals, particularly wild boars. Its geography is currently limited mainly to Japan, and information on its pathogenicity in humans is limited. Genotype 6 is largely similar to genotype 5, with cases also detected in wild boars in Japan. The newest of the identified genotypes are genotype 7 and 8, which were detected in dromedaries and Bactrian camels, but a link to human infections has not yet been established (Ejaz et al., 2024).

It is also important to mention the high frequency of hepatitis E mutations in its genome. According to estimates, the mutation rate may be a little lower than 1.40×10^{-3} base substitutions per site per year. This value is comparable to the mutation rate of hepatitis C, for example. There are several reasons for these mutations. The first is the absence of corrective mechanisms. The viral RNA-dependent RNA polymerase of hepatitis E does not have a proofreading activity. This leads to the accumulation of mutations during replication. Such feature can lead to the presence of quasi-species at within-host level. These, and the virus population at epidemiological scale, is subjected to the action of the immune pressure. from the host. The virus is forced to adapt to immune responses, which can lead to changes as it attempts to escape recognition by antibodies or immune cells. An additional driving force can be the uncontrolled use of antiviral drugs, such as ribavirin. It causes mutations in the hepatitis E genome, which potentially leads to the formation of new and different viral populations.. However, it should be noted that not all mutations lead to the formation of viable viruses. All of the above mutations in the virus genome can affect both pathogenesis, influencing the severity of the disease, antigenicity, and resistance to antiviral drugs (Hoang van Tong et al., 2016).

1.1.2. Evolutionary and genetic feature

According to recent studies, the evolution of hepatitis E began around 6,000 to 6,800 years ago. This coincides with the period when humans domesticated pigs. This process created the ideal conditions for the virus to adapt to its new host. At the same time, the last common ancestor of all hepatitis E genotypes existed between 536 and 1344 years ago. Apparently, this ancestor gave rise to anthropotropic and zoonotic forms of hepatitis E, which later evolved into genotypes 1, 2, and 3, 4, respectively. The dynamics of the hepatitis E population indicate that genotypes 1, 3, and 4 underwent significant expansion during the 20th century. Genotype 1 increased in size of the infected population approximately 30–35 years ago. Genotypes 3 and 4 experienced an increase in population size from the late 19th century to approximately 1940–1945, with genotype 3 experiencing an additional rapid expansion until approximately 1960. The effective population size for both genotype 3 and genotype 4 rapidly declined to pre-expansion levels beginning around 1990 (Purdy et al., 2010).

Let's take a look at the molecular side of hepatitis E evolution. Its genetic material has an interesting feature called codon bias. This is a phenomenon in which certain codons are used more often than others to encode the same amino acid. This is a phenomenon in which certain codons are used more often than others to encode the same amino acid. Since codon bias features host species as well, HEV codon bias might be indicative of pressures pointing at host adaptation. In fact, hepatitis E codon bias is caused by both mutational pressure and natural selection. For example, in some open reading frames of hepatitis E, there is a predominance of codons ending in U and C. This may be due to forced adaptation to the conditions for replication in the host cells of the virus. It has also been proven that ORF1 shows the least bias. ORF 2 and ORF3, on the other hand, show a more pronounced preference for certain codons. This fact may indicate that different ORFs may undergo different evolutionary changes, which in turn affects their molecular evolution and ability to adapt to different hosts (Baha et al., 2019).

1.2. Zoonotic transmission pathways and animal reservoirs of HEV

As already mentioned, hepatitis E virus is a unique pathogen that possesses both anthroponotic and zoonotic transmission mechanisms. The most epidemiologically significant genotypes are HEV-3 and HEV-4, which can be transmitted from animals to humans. These genotypes are found in a variety of mammals, including domestic pigs, wild boars, deer, rabbits, camels, and rodents, creating an extensive zoonotic reservoir of the virus in both wild and agricultural environments (Lhomme et al., 2025; Meng, 2013; Abravanel et al., 2024).

One of the most documented routes of HEV transmission from animals to humans is the foodborne pathway (Figure 1). Numerous epidemiological and molecular studies confirm a link between the consumption of raw or undercooked meat, particularly the liver of pigs, wild boars, and deer, and HEV infection (Tei et al., 2003; Bendall et al., 2012). In European countries such as France and Germany, outbreaks of HEV have been associated with the consumption of traditional products, including dry-cured sausages made from pig liver (figatelli) (Boutrouille et al., 2007). In Japan, cases of acute hepatitis E have also been linked to the consumption of raw wild boar and venison (Tei et al., 2003). Additionally, HEV has been detected in unpasteurized camel milk, indicating the potential for virus transmission through dairy products (Lee et al., 2016).

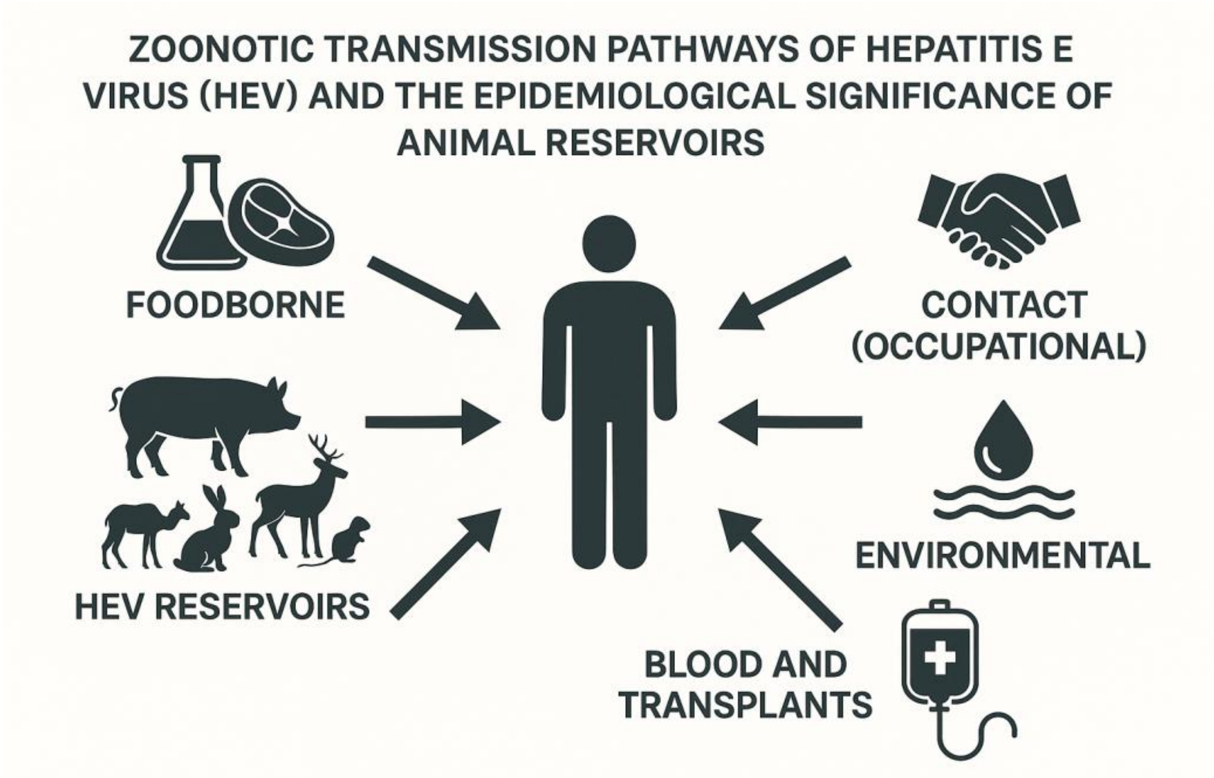


Figure 1. Zoonotic transmission pathways of hepatitis E virus (HEV)

Another significant transmission route is the contact pathway. Workers in pig farming, the meat processing industry, abattoirs, veterinarians, and hunters are at high risk of infection due to contact with infected animals or their biological fluids. Studies indicate that the seroprevalence of HEV among farm workers is significantly higher than in the general population, especially in countries with high pork consumption (Matsuda et al., 2003; Adlhoch et al., 2021).

In the context of inadequate sanitation infrastructure, environmental transmission of HEV may occur through the contamination of the environment with feces from infected animals. HEV is resilient in aquatic environments and can maintain infectivity in wastewater, soil, and on surfaces, facilitating indirect transmission to humans via contaminated water, agricultural products, or contact with contaminated environments (Miura et al., 2020; EFSA, 2017).

It is also important to note blood and transplant transmission, especially considering that a significant portion of HEV infections are asymptomatic. Cases of infection through donated blood containing HEV-3 have been reported in the UK, France, and the Netherlands, likely originating from donors infected via zoonotic exposure(Hewitt et al., 2014).

A wide range of species capable of serving as HEV reservoirs is confirmed by both field observations and experimental models. Domestic pigs are recognized as the primary zoonotic reservoir, as the virus circulates actively in their populations and these animals often remain asymptomatic carriers (Krog et al., 2013). High prevalence of HEV is also observed in wild boars and deer, particularly in areas with active hunting. Rabbits are carriers of a specific subtype, HEV-3ra, that is capable of infecting humans (Izopet et al., 2012).

1.3. Host-pathogen interactions

During their evolution, RNA-genome viruses have developed various complex strategies that allow them to survive in the host organism, even when conditions are unfavorable. These strategies not only ensure the viruses' resistance to physical and chemical stresses but also allow them to actively exploit the host's cellular mechanisms for their replication and survival.

A crucial aspect of the RNA virus life cycle is the ability to exploit host cell resources to replicate their genetic material. To achieve this, viruses employ a variety of molecular strategies. For instance, they interact with host translation factors that are essential for the initiation of protein synthesis. A well-documented example is provided by poliovirus, hepatitis A virus, and hepatitis C virus, which utilize internal ribosome entry sites (IRES) dependent on host translation factors to selectively initiate viral protein synthesis. Moreover, other viruses, including hepatitis C virus and herpes simplex virus, adopt mechanisms involving the phosphorylation of translation initiation factors, thereby enhancing the synthesis of their own proteins. Through such strategies, viruses manipulate host cellular pathways, reprogram translation regulation, and ultimately optimize conditions for their replication (Alcami & Koszinowski, 2000).

This phenomenon is true for HEV also, showinga complex, ongoing process of interaction between the hepatitis E virus and the host's cellular factors throughout the virus's life cycle. Clearly, the first and necessary step involves the use of certain cellular receptors to attach to and get into the host's cells. Hepatitis E, which is the focus of this paper, binds to heparan sulfate proteoglycans and the sialoprotein receptor, which helps it get into cells.

At the viral replication level, the hepatitis E virus interacts with different host proteins. For example, translation initiation factors eIF4A and RACK1. They are necessary for successful virus reproduction. These interactions not only ensure viral RNA replication, but also allow the virus to manipulate cellular signals in the host organism. This can lead to changes in the immune response and disruption of cellular processes. For example, viral proteins can interfere with host cell signaling pathways and anti-viral systems, helping the virus evade the immune response (Rein et al., 2021).

More in detail, after entering the host organism, viral particles bind to liver cells receptors, such as heparan sulfate proteoglycans and virus initiates a process in which the cell membrane envelops the viral particle, forming a vesicle containing the virus. The vesicle with the virus sinks into the cytoplasm of the cell forming an endosome, which can then merge with other intracellular organelles like lysosomes. The process leading to the release of the viral genome from the endosome into the cytoplasm has not been precisely established. However, as a rule, this is associated with a change in pH and protease activity, which allows the virus to release its RNA.

The next stage in the virus life cycle is the release of viral RNA into the cytoplasm. There, it is used as a template for the synthesis of viral proteins. This process begins with the translation of the non-structural polypeptide pORF1. pORF1 contains several functional domains, including a methyltransferase, a helicase, and an RNA-dependent RNA polymerase. The mechanisms of its processing are not well understood. What can be said at this point is that processing may depend on guest proteases, such as thrombin and factor Xa, which are involved in blood coagulation. Their activities are essential for replication of hepatitis E virus and are possibly implicated in ORF1 polyprotein processing

Once produced, the RdRp initiates the transcription of the viral genomic RNA. This intermediate RNA acts as a template for producing new positive-sense viral genome copies and protein translation also occurs. Viral encapsidation and assembly are facilitated by the capsid protein's interaction with a 76-nucleotide region specific to the 5' end of the HEV genome. Following assembly, the viral progeny are transported via multivesicular bodies and released through the cellular exosomal pathway. This allows them to avoid the host's immune response.

The host's immune responseplays a key role in controlling and eliminating viral infections. It includes both innate and adaptive immune mechanisms that interact with each other to protect the body. When the hepatitis E virus enters the body, it activates the innate immune response. Next, macrophages and dendritic cells recognize viral particles through pattern recognition receptors like toll-like receptors. After the virus is recognized, pro-inflammatory cytokines - as interleukin-6 and interferon-alpha - are released. Their role is to intensify inflammation and attract other immune cells.

In turn, adaptive immunity, which develops in response to infection, provides protection against reinfection. In the case of hepatitis E, T lymphocytes are activated, which recognize infected cells. CD8+ cytotoxic T cells play a particularly important role. They destroy cells containing the virus. T cells are activated by antigen-presenting cells. These can be dendritic cells, which capture viral antigens and present them on their MHC molecules.

The next important step is the production of antibodies. The activation of B lymphocytes leads to the production of specific IgM and IgG antibodies against the virus. They help neutralize viral particles and prevent reinfection. After successfully fighting the infection, some B cells turn into memory cells, which are capable of providing long-term protection and the ability to respond quickly to reinfection.

During the activation of T and B cells, various cytokines are released that regulate the immune response, enhance inflammation, and attract additional immune cells to the site of inflammation. An example of such a cytokine is interferon-γ. It plays an important role in the activation of macrophages and other cells of the innate immune system, which enhances the overall immune response.

In healthy people without chronic diseases, the adaptive immune response allows them to successfully control hepatitis E infection. This usually leads to a complete recovery without the hepatitis progressing to the acute phase. People with weakened immune systems or liver conditions are at risk in of severe disease progression.

1.4. Environmental surveillance of HEV

Hepatitis E, as already mentioned, is a zoonotic disease that is widespread in the environment and poses a serious global health problem. This is why environmental monitoring is necessary. This comprehensive process includes assessing the prevalence of the virus in natural water bodies, drinking water, agricultural products, and animal reservoirs, which allows for the timely identification of sources of infection and prevention of disease outbreaks. The main source of environmental contamination with HEV is water contaminated with feces, both drinking water and water used for domestic purposes. In countries with poor sanitation, the lack of proper wastewater treatment contributes to the spread of the virus through waterways, including rivers, lakes, and municipal systems.

The hepatitis E virus is resistant to extreme environmental conditions. It can survive drying, ultraviolet radiation, and chemical treatment, which increases its survivability and potential for transmission through water and food. Even in economically developed countries, such as those in the European Union, despite rare cases of waterborne outbreaks, significant circulation of HEV has been detected in wastewater and surface water. This indicates the need for regular monitoring of water resources.

Methods for concentrating viruses from various environmental matrices play an important role in HEV environmental monitoring.. Some of the main methods are membrane filtration, ultracentrifugation, and precipitation using polyethylene glycol. Membrane filtration is effective for large volumes of water, while ultracentrifugation provides high virus concentration but requires expensive equipment. Precipitation using polyethylene glycol is more affordable but labor-intensive and less effective. The choice of method depends on the type of sample and research conditions. The main limiting factor in this situation is the lack of standardization of methods for concentrating and detecting hepatitis E. This also makes it difficult to compare data and form a clear picture of the spread of the virus in nature. More traditional RNA extraction techniques are typically performed on food matrices, animal tissues or clinical specimens, although proper extraction method validation is fundamental for reliable results.

Molecular biological methods such as RT-qPCR and new isothermal methods are widely used to detect HEV in environmental samples. New methods include, for example, RT-RPA. However, molecular methods only detect viral RNA and cannot determine the infectious potential of the virus. This is a serious limitation when assessing the real threat to public health (Elois et al., 2025).

1.4.1. Foodborne HEV outbreaks in Italy

Although hepatitis E is often not regarded as a major threat in high-income countries, its relevance is frequently underestimated, particularly with respect to its environmental presence, circulation in animal populations, and clinical impact. Three studies conducted in Italy provide illustrative examples of these issues.

Let's consider a real case of environmental surveillance of hepatitis E. In a study covering a nine-year period from 2011 to 2019, 1,374 wastewater samples from 48 treatment plants located in all 20 regions of Italy were analyzed. The use of molecular methods, including PCR and sequencing, revealed the presence of hepatitis E virus RNA in 5.4% of samples. The highest detection rate, 7.1%, was observed in the central regions of the country, such as Lazio and Abruzzo. This information is consistent with clinical epidemiology and serological studies, which indicate a high level of virus circulation in these regions.

Genotypic analysis revealed the dominance of genotype 3. It was observed in 76% of cases. This is characteristic of autochthonous infections associated with zoonotic transmission from domestic and wild pigs. Genotype 1, mainly brought in by infected tourists from Asia and Africa, was less common and mainly occurred in the early years of observation. Quantitative analysis showed the presence of the virus in wastewater in concentrations ranging from several dozen to tens of thousands of genomic copies per liter. This fact indicates a constant but non-epidemic level of HEV prevalence in the Italian population. This situation should not be ignored.

The results of this study highlight the need to integrate environmental and clinical monitoring of the virus for effective management of hepatitis E-related risks (Iaconelli et al., 2020).

Between June and December 2019, there was a surge in hepatitis E virus infections in central Italy, in the regions of Abruzzo and Lazio. A total of 47 cases of infection were reported. This figure is significantly higher than in previous years. Among those who fell ill, 44 people were hospitalized with acute hepatitis, and three cases were asymptomatic infections detected in transplanted organs. Men predominated among those who fell ill, and the average age of patients was 63 years.

Epidemiological studies found that most patients reported consuming pork products. These products included raw or undercooked sausages. This fact helped epidemiologists link the outbreak to a potential source of infection. This source turned out to be meat products containing the hepatitis E virus. A distinctive feature of this outbreak was that among the infected patients there were cases with more severe symptoms than usually observed, including high levels of transaminases and jaundice.

Analysis of virus sequences obtained from 35 of the 47 cases revealed the presence of several virus clusters, including subtypes 3e and 3f. The main clusters A and B were found in both areas, while cluster C, which also contained 11 sequences, showed a high degree of relatedness to other strains but was geographically more widespread. These data suggest that the outbreak may have been caused by new strains that had not previously circulated in this region.

After the outbreak was detected, local health authorities launched an active investigation, including food sampling and supply chain tracing. However, most of the suspect food batches were no longer available for testing. This made it difficult to identify the source of the infection. Fortunately, the outbreak resolution was spontaneous: two months after the last reported case, no new infections were observed (Garbuglia et al., 2021).

In 2021, a study was conducted on the spread of hepatitis E virus in wild boars in Tuscany. As part of the study, liver tissue samples and carcass swabs were collected from wild boars that underwent sanitary inspection in butcher shops after hunting. The results showed that 12% of liver samples and 6% of swabs contained hepatitis E virus RNA. This indicates that the virus is actively circulating among the wild boar population in Tuscany. It is important to note that some swabs tested positive even though liver samples from the same animals were negative. This indicates possible cases of cross-contamination during slaughter (Forzan et al., 2021).

Together, these studies emphasize that HEV circulation in Italy is not negligible. They highlight the need for strict compliance with hygiene standards in slaughterhouses and food processing plants, as well as the importance of raising awareness among hunters, meat industry workers, and consumers. Ensuring proper cooking of pork and game products and monitoring raw materials remain critical preventive measures against HEV transmission.

2. AIM OF THE THESIS

The aim of the thesis is to investigate the use of neural networks to predict the host source of the hepatitis E virus based on nucleotide data using a bioinformatic approach. The work itself is aimed to determine whether computational methods can effectively solve host identification problems. The results of the study are important for food epidemiology, as accurate host identification can significantly influence the understanding of virus transmission dynamics and evolution. Moreover, if successful, this would testify to the presence of a viable computational framework that enhances predictive accuracy in virology, paving the way for improved public health responses and interventions. This research could lead to more effective monitoring and control strategies for hepatitis E outbreaks, ultimately contributing to better management of foodborne viral infections.

In recent years, computational biology has had a significant impact on virology and pathogen monitoring. Deep learning models are of particular interest. They are important for their ability to process large amounts of biological data and identify patterns. This puts them at an advantage over traditional methods. As part of this master's thesis, a training model of a simple single-layer neural network was developed. It is designed to classify viral sequences depending on the host species. The principle of operation of this model is based on the fact that it uses numerical representations of nucleotide sequences obtained by encoding k-mer frequencies. This is a method that converts sequence data into a format that is both informative and efficient.

The dataset used to train this model consists of carefully selected hepatitis E genomes. Each genome is accompanied by annotated host metadata. This careful selection ensures that the data is not only relevant but also formatted for machine reading. The latter significantly increases the model's learning potential.

During the selection and processing of metadata, preprocessing was performed to download, clean, and encode viral sequences extracted from FASTA files. The annotation of the sequence with the verified host information was manually curated to increase the reliability of the training dataset. The classification efficiency of this model was carefully evaluated using metrics such as accuracy, precision, and recall. These criteria allowed to conduct the neural network's accuracy assessment.

In conclusion, the aim of this thesis is to evaluate the possibility of integrating neural networks into the field of viral genome analysis. This research can bring about positive changes in food safety control. Also in the field of epidemiology, accurate predictions of host origin can serve as the basis for public health strategies and measures to respond to disease outbreaks. In addition, understanding of the dynamics of transmission between hosts and viruses can be crucial in zoonotic risk assessment and preventing potential outbreaks.

3. MATERIALS AND METHODS

3.1. Data Sources and Preprocessing

The present thesis is grounded in the analysis of hepatitis E virus genomic sequences and rigorously curated host-associated metadata. In this case, data preprocessing is a critically important step in developing an accurate host prediction model. The initial dataset consisted of complete hepatitis E virus genome sequences obtained in FASTA format from publicly available databases, with a primary focus on the NCBI GenBank repository. Sequence download and extraction was achieved by writing Python code to collect the data from mentioned open source. Further work with the metadata annotation was done manually. A major challenge was that a significant amount of data was entered into the database either incorrectly formatted or incomplete. Sequences for the study were selected based on several criteria. These criteria included completeness, annotation quality, and the availability of accompanying host information. To ensure clarity and traceability at all stages of preprocessing, each sequence was named according to its GenBank accession number.

Next, an edited file with characteristics in Microsoft Excel format was manually created. It contained two important columns: access and host. This file was named "traits_updated_with_all_human_hosts.xlsx." This file was carefully edited to correct inconsistencies in host nomenclature, standardize organism names. Only entries that contained clear, biologically meaningful information about the host were retained for further analysis.

Another Python script was written to merge host metadata with FASTA sequence files. This merging process involved matching access numbers from an Excel file with file names in the FASTA directory. Any sequences without corresponding metadata were automatically excluded. This ensured that the final dataset consisted only in verified and labeled records. These actions made it possible to obtain a clean and reliable dataset, which is necessary for training a neural network model (Caliskan et al., 2023).

In terms of genotypes, the study concentrated on genotypes 3 and 4, focusing on the critical interface of zoonosis. This approach aimed to uncover genetic signals in the viral genome that may indicate host preferences or tropisms. By limiting the dataset to genotypes 3 and 4, the study avoided artificially inflated performance due to genotype-host correlations, as genotype 1 is exclusively associated with humans (Wang and Meng, 2021). This focus allowed the model to discriminate host origin based on more subtle genomic signals within zoonotic genotypes. Additionally, a phylogenetic tree was constructed to illustrate how the viral sequences cluster based on their host species, revealing potential transmission routes between animals and humans over time. By utilizing the same genotypes 3 and 4 in both the neural network and the phylogenetic analysis, the tree serves to validate whether the host predictions made by the model align with real genetic relationships among viruses from different hosts ((Reiman et al., 2020).

3.2. Machine learning approach overview

In this stage, a thorough analysis was conducted to evaluate the ability of the neural network to classify hepatitis E virus (HEV) sequences based on host species. The main objective was to assess the performance of the model, its capacity to generalize, and the biological implications of the findings.

To facilitate this analysis, the dataset, consisting of 448 sequences, was divided into two subsets: the training set, which included 381 sequences (85%), and the test set, comprising 67 sequences (15%). During model development, cross-validation was applied within the training set to optimize performance and reduce overfitting (Matuszewski and Sintorn, 2021). A balanced representation of classes was ensured in this subset to minimize bias during training and to capture the necessary diversity across host species. Finally, the test dataset, which remained completely unseen throughout training and validation, was used solely to evaluate the final performance of the model, providing an unbiased assessment of its classification capabilities.

During the training process, the cross-entropy loss was monitored, providing a quantitative evaluation of the discrepancy between predicted and actual labels. This loss function was selected due to its efficacy in multi-class classification, as it imposes greater penalties on incorrect predictions based on the degree of deviation from the correct class probability.

Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, which were calculated on the test dataset following the convergence of the training process. As the test set is entirely separate from the data used to train and tune the model, it ensures that the results reflect genuine predictive power rather than mere memorization of the training or validation samples. Accuracy was defined as the ratio of correct predictions to the total number of samples, serving as a general indicator of model efficacy. Precision and recall were examined for each host class, enabling the identification of potential class imbalances and instances of overfitting to dominant categories (Erickson and Kitamura, 2021).

A confusion matrix was generated to illustrate the classification behavior of the model, highlighting true positive rates and misclassification patterns across host classes. This allowed for the identification of specific hosts that were frequently misclassified, offering insights into biological similarities or inconsistencies in the training data annotations. Additionally, recall curves were produced to assess the sensitivity of the model across individual classes, which is critical in scenarios where false negatives could have significant consequences, such as in zoonotic spillover risks.

All statistical analyses and visualizations were executed using Python libraries, including scikit-learn, NumPy, and Matplotlib.

3.3. k-mer Encoding

To make biological sequence data comprehensible for machine learning algorithms, it is essential to convert raw nucleotide strings into a numerical format that retains biologically significant patterns. K-mer encoding was applied in this study. It is a widely recognized technique in computational biology that helps achieving this transformation. A k-mer is defined as a subsequence of length k extracted from a longer DNA or RNA sequence by sliding a window across it, moving one nucleotide at a time (Moeckel et al., 2024). The frequency distribution of these subsequences reveals local structural features and base composition. These elements can provide valuable insights into the evolutionary origins and host specificity of viral genomes.

In the present study, k=6 was selected thus, as seen in Formula (1), 4096 potential k-mer combinations were present. We selected from the four standard nucleotides: A, T, C, and G.

$$N=|\Sigma|^k \qquad (1)$$

$N=4^6=4096$

Each HEV genome was then converted into a vector representing these 4096 features, where each feature indicated the normalized frequency of a specific 6-mer. This normalization process was very important. It ensured that variations in genome length did not skew the input data. It gives an opportunity for a reliable comparisons across sequences of different sizes. The choice of $k = 6$ was supported by previous research that demonstrated that 6-mers effectively capture taxonomic and functional differences among viral strains, all while keeping the feature space manageable (Zhang, Q. et al. 2017).

The implementation of k-mer encoding was carried out in Python, utilizing Biopython for parsing FASTA files and NumPy for efficient vector operations. Initially, sequences were cleaned to eliminate non-standard characters and unclear nucleotide codes. It was ensured that only valid A, T, C, and G substrings were included. The script was created to analyze each genome, extracting all overlapping 6-mers, counting their occurrences, and normalizing these counts based on the total number of k-mers in that particular sequence. The result was a feature matrix where each row corresponded to a viral genome, and each column represented a specific 6-mer, thereby constructing a high-dimensional representation of the viral sequence dataset.

Straightforward and interpretable nature of k-mer encoding makes it a compelling alternative to more complex approaches. For example, alignment-based methods or deep sequence embeddings. What is more, k-mer encoding is alignment-free (Ren et al., 2018). It allows it to scale more effectively with larger datasets.

In conclusion, the k-mer encoding step acted as a pivotal link between biological sequence data and machine learning models. It enabled the preparation of critical information about genomic content by transforming sequences into an analyzable format suitable for statistical analysis and classification. This transformation set the stage for the neural network model to uncover patterns of host adaptation embedded within the viral genome.

3.4. Model Training and Evaluation

The training of the neural network model was conducted on a meticulously curated training dataset derived from hepatitis E nucleotide sequences. Each sequence was annotated with a verified host species. These sequences underwent preprocessing utilizing the k-mer encoding method previously detailed. The resulting feature vectors were then aligned with their corresponding class labels, thereby constructing a comprehensive training dataset.

Also one-hot encoding was applied on the host labels before the training step. It transformed each categorical class into a binary vector. This format is requisite for the softmax output layer of the network and is critical for accurately computing the categorical cross-entropy loss. Furthermore, the input features were standardized to ensure consistent scaling, which expedited convergence and stabilized the learning process.

Training was executed across multiple epochs, usually ranging from 50 to 100, with batch sizes fluctuating between 16 and 64, contingent upon the available computational resources. The Adam optimizer was utilized to update the model weights, chosen for its efficacy in managing sparse gradients and noisy data distributions (Yang, 2024). A training set was employed to monitor the model's performance at the conclusion of each epoch. Throughout the training epoch, training accuracy, along with loss metrics, was meticulously tracked and plotted. These visualizations facilitated an in-depth examination of the learning dynamics. A characteristic sign of effective training is a consistently decreasing loss curve for the training set, accompanied by a corresponding increase in accuracy.

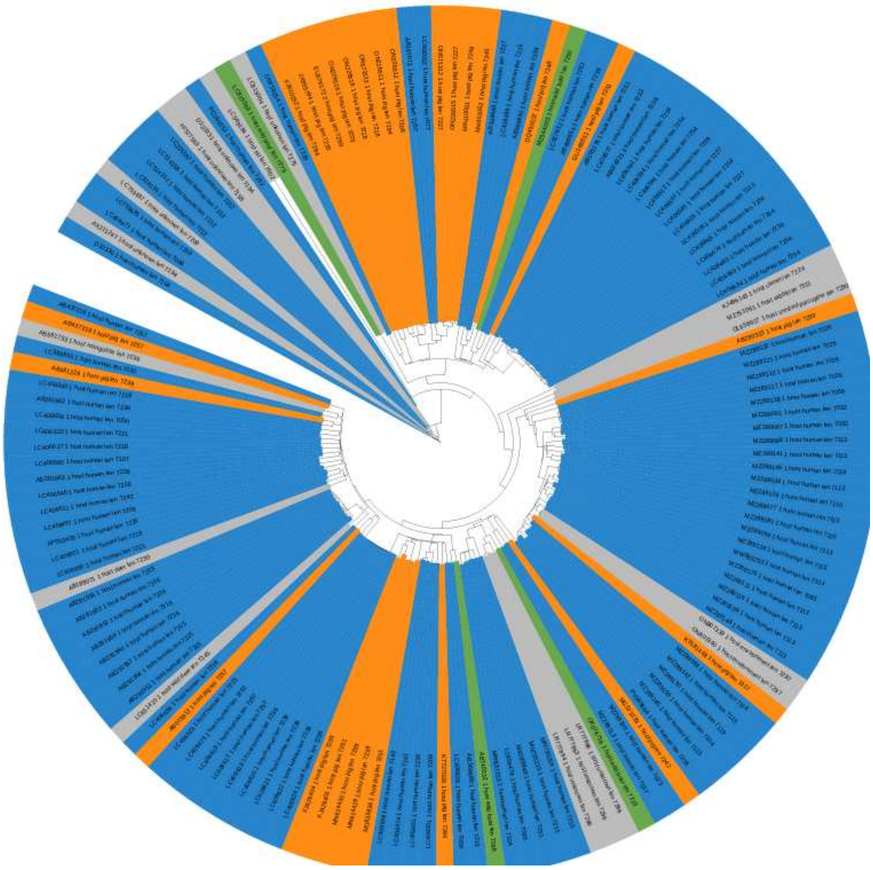
Upon the completion of training, the final model was rigorously evaluated using a suite of metrics - precision, recall and a confusion matrix - calculated on the test dataset comprising 67 sequences (15% of the total 448).

To enhance interpretability, a visual analysis of the model's predictions was conducted. Specifically, confusion matrices were generated to illustrate the frequency with which one host was misclassified as another. These matrices not only highlighted the most distinguishable host classes but also identified those that presented challenges due to sequence similarity or insufficient representation in the training dataset.

4. RESULTS AND DISCUSSION

4.1. Zoonotic risk assessment of HEV through phylogenetic tree

The preliminary phylogenetic tree (Figure 2) performed on genotypes 3 and 4 confirmed the zoonotic potential of these viruses, as strains collected from different host species were interspersed throughout the tree, suggesting frequent cross-species transmission. At the same time, the presence of species-specific clusters might also indicate a certain degree of host adaptation.



Host	Color
human	
pig	
wild_boar	
unknown / other	

Figure 2. Phylogenetic tree of Hepatitis E genotypes 3 and 4. Different host species have been color coded

More in detail, one of the most striking observations is the recurrent nesting of human isolates (blue) within pig-dominated clusters (orange). For example, human sequences such as LC406560 and LC406658 are situated directly alongside pig isolates, including FJ426404 and MN614430, with minimal branch lengths separating them. This close relationship suggests active zoonotic exchange, supporting the notion of foodborne transmission through pork products. Moreover, occupational exposure among individuals such as farmers, butchers, and abattoir workers further amplifies this risk. The consistent repetition of these clusters across the phylogenetic tree indicates that spillover is not an isolated event but a persistent characteristic of HEV genotype 3 epidemiology.

In addition to pig associations, the phylogenetic tree reveals significant clustering of human isolates with wild boar sequences (green). For instance, the human sequence MZ289113 closely aligns with wild boar sequences AB740232 and OK076716. This arrangement underscores the ecological overlap between domestic pigs and wild boars, facilitating the circulation of HEV among both managed and wild populations. Hunting practices and the consumption of wild game meat, particularly noted in regions such as Europe and Japan, further heighten the risk for humans, positioning wild boars as critical bridge hosts in the zoonotic transmission landscape.

The analysis also uncovers environmental or unclassified isolates (grey) embedded within suid-rich clades, such as ON807339 and ON807340. Their positioning provides molecular evidence of viral shedding into shared ecological resources, including water and shellfish, which can act as indirect transmission pathways. This observation broadens the zoonotic risk profile, extending it beyond occupational or dietary groups to encompass the general population, as evidenced by outbreaks linked to contaminated water and seafood.

The phylogenetic tree highlights the presence of secondary hosts as well. For example, the rabbit sequence MZ751061 is categorized within genotype 3 but remains relatively distant from human and pig clusters. This suggests a lower, yet non-negligible, spillover potential, especially in contexts where rabbit meat or liver is consumed. Other wildlife sequences, such as those from deer (AB189071) and mongoose (AB591733), are positioned more distantly, indicating a broader host diversity for HEV while suggesting lower immediate zoonotic risk.

Geographic structuring is also evident in the phylogenetic analysis. Genotype 3, prevalent in Europe and Japan, displays dense clusters of human-pig intermixing (blue/orange). Conversely, genotype 4, more common in East and Southeast Asia, shows greater involvement of wild boars (green) alongside human and pig sequences, reflecting local ecological interfaces and dietary practices unique to these regions.

From a risk assessment perspective, the color-coded phylogeny elucidates three critical transmission pathways:

1. Foodborne Route: Facilitated by repeated human-pig clustering, emphasizing the role of pork and wild game in transmission.
2. Occupational Route: Highlighted by the frequency of human sequences within pig clades, which points to the risks faced by farmers, veterinarians, and abattoir workers.
3. Environmental Route: Evidenced by the embedding of environmental isolates in suid clusters, indicating potential contamination of water and shellfish.

In summary, the phylogenetic tree - interpreted alongside the provided host legend - demonstrates the multi-layered zoonotic potential of HEV genotypes 3 and 4. The consistent clustering of human sequences with pig and wild boar isolates, the incorporation of environmental samples, and the identification of secondary reservoirs confirm the intricate nature of HEV transmission dynamics.

4.2. Dataset composition

This study used a carefully compiled dataset that included complete or nearly complete genomes of the hepatitis E virus. Each genome was assigned a reliably verified host label. The main hosts were humans, pigs, wild boars, rabbits, and several other wild animals. A small number of sequences were annotated as originating from environmental samples, but these were treated with caution. In these cases, no specific metadata was available to determine the source context, and as such, these samples were excluded from model training and evaluation to avoid introducing noise or biologically ambiguous patterns. Additionally, sequences from experimentally infected animals were manually screened and excluded, as they do not reflect natural host tropism and could bias the classification model. To ensure data integrity, genome sequences were identified by their accession number and then merged with a manually refined feature table containing two key columns: accession number and host. During the refinement process, records with missing, inconsistent, or ambiguous host information were systematically removed. This careful exclusion helped maintain the reliability of the dataset and prevent potential errors in further analysis.

To prevent distortion of performance indicators, sequences that were nearly duplicates were significantly reduced. This applies particularly to sequences originating from the same outbreak or laboratory source. The reduction in significance occurred during the data separation stage. This approach minimized the risk of the model becoming biased due to highly similar genomes. Thus, it helped prevent overfitting on repetitive patterns that cannot be generalized.

The distribution of host classes in the collected dataset is significantly uneven, as shown in Figure 3. Humans and pigs dominate the data, accounting for the majority of sequences—267 and 175, respectively. Although rat-derived sequences appear frequently, a substantial proportion of these are likely associated with experimental infections or rat cell line passages, rather than representing true natural hosts. To maintain biological validity, a manual screening process was implemented to identify and exclude sequences derived from experimental settings or ambiguous sources**,** especially where metadata indicated cell culture or artificial infection. Wild boars, rabbits and other wildlife species are represented in much smaller numbers. For example, there are only 19 wild boar sequences, and other hosts such as monkeys, camels, and mice occur in single or only one instance. This marked imbalance reflects a general trend in public genomic repositories, where selection efforts are most often focused on clinical and agricultural issues rather than on achieving an epidemiologically balanced dataset.

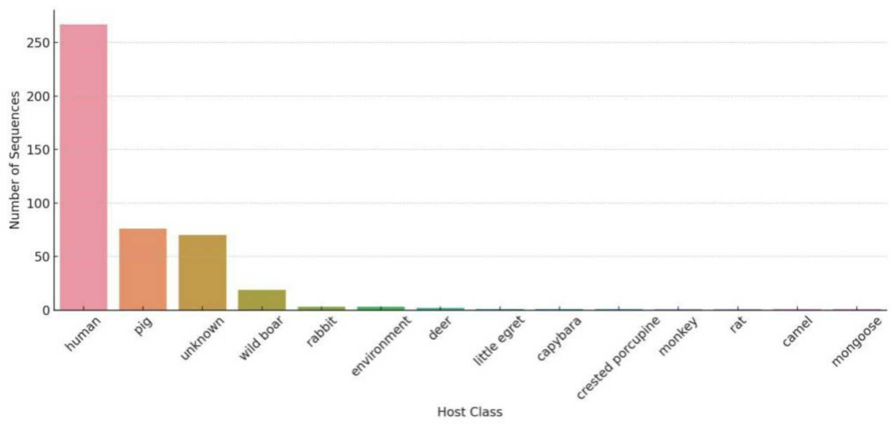


Figure 3. Number of HEV Sequences per Host Category

This skewed representation presents a significant challenge for predictive modeling. Discriminative algorithms inherently tend to favor majority classes, often at the expense of accurately identifying minority categories. To mitigate this bias, a stratified data-splitting strategy was employed, ensuring that train, validation, and test subsets preserved the original class proportions. Moreover, during model optimization, class weighting was incorporated so that errors made on rare host classes carried more influence in the loss function. This technique encourages the model to pay closer attention to under-represented hosts, thereby enhancing its ability to generalize beyond the dominant classes.

It is crucial to recognize that this imbalance is not simply a statistical inconvenience, it has real-world implications for zoonotic risk assessment. The under-representation of certain wildlife hosts may obscure important transmission pathways, potentially leading to gaps in public health preparedness. As such, the painstaking efforts to account for these imbalances in modeling are not merely technical adjustments but vital steps toward a more comprehensive understanding of HEV ecology. It can almost be frustrating to see how the dataset reflects the biases of human priorities rather than the complex realities of viral ecology, yet these challenges underscore the importance of sophisticated modeling strategies.

All in all, this hepatitis E dataset reflects a complex interplay between biological diversity and sampling bias. The dominance of human and pig sequences stands in stark contrast to the sparse representation of other hosts, shaping the behavior and limitations of predictive models. Through careful data research, stratified splitting, and class-weighted optimization, the study strives to confront these challenges head-on, aiming to build more robust and precise zoonotic risk predictions.

4.3. Neural Network Nodel Architecture

The predictive model for host classification utilized a fully connected feed-forward neural network. This architecture was selected due to its simplicity and demonstrated effectiveness in bioinformatics tasks that involve structured feature vectors. For example, such vectors like k-mer profiles. Model was implemented by the Keras API, which is built on TensorFlow.

As mentioned above, the network's input consisted of 4096-dimensional vectors that represented normalized 6-mer frequencies from each viral genome. To effectively handle this high-dimensional input, the model began with an input layer sized at 4096, directly reflecting the feature space defined by all possible 6-mer combinations.

Next, the architecture incorporated two hidden dense layers. The first dense layer contained 512 neurons and employed the Rectified Linear Unit activation function, introducing non-linearity while avoiding issues associated with vanishing gradients. To overcome overfitting and improve generalization to unseen data, a dropout layer with a dropout rate of 0.3 was applied. This layer randomly deactivates 30% of neurons during training. It pushes the network to learn more clean-cut and less redundant. Following this, a second hidden layer with 128 neurons also utilized the ReLU activation and included another dropout layer with the same configuration.

The model concluded with a softmax output layer, the size of which matched the number of unique host classes in the dataset. The softmax function is to transform the raw output scores into probabilities that sum to one. It enables the model to make multi-class predictions. Each output neuron corresponded to a potential host class, with the neuron yielding the highest probability indicating the predicted host species.

To provide a clearer understanding of th e model's structure and flow, refer to the following diagram (Figure 4) :

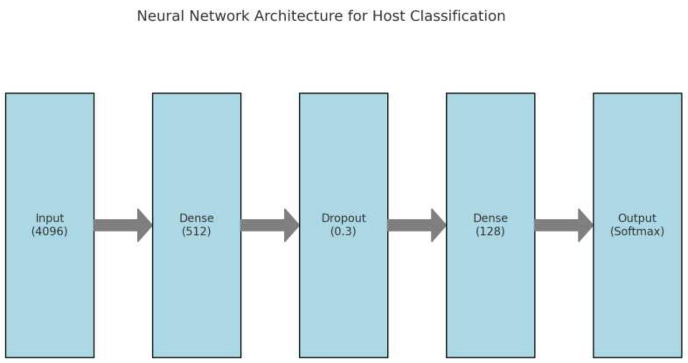


Figure 4. Schematic depiction of the Neural Network architecture implemented in the present work.

This architecture was intentionally designed to strike a balance between computational efficiency and the ability to learn meaningful patterns from genomic input data. Its moderate depth and dropout regularization make it well-suited for the volume of available training data, reducing the risk of overfitting while achieving high classification performance. This design formed the backbone of the hepatitis E host prediction system. Also it was iteratively fine-tuned based on the results obtained from experimental training.

4.4. Training dynamics and Cross-Validation Learning Curves

To evaluate the training dynamics and generalization capabilities of the neural network classifier designed for predicting Hepatitis E Virus (HEV) hosts, a stratified 5-fold cross-validation approach was utilized. This methodology was essential to ensure that the distribution of classes remained consistent across each fold, a factor that was particularly significant due to the inherent host imbalance present in the dataset. During the execution of this process, four folds were employed for training the model while the remaining fold served as the validation set. This sequence was repeated across all five folds, generating a comprehensive estimate of the model's performance.

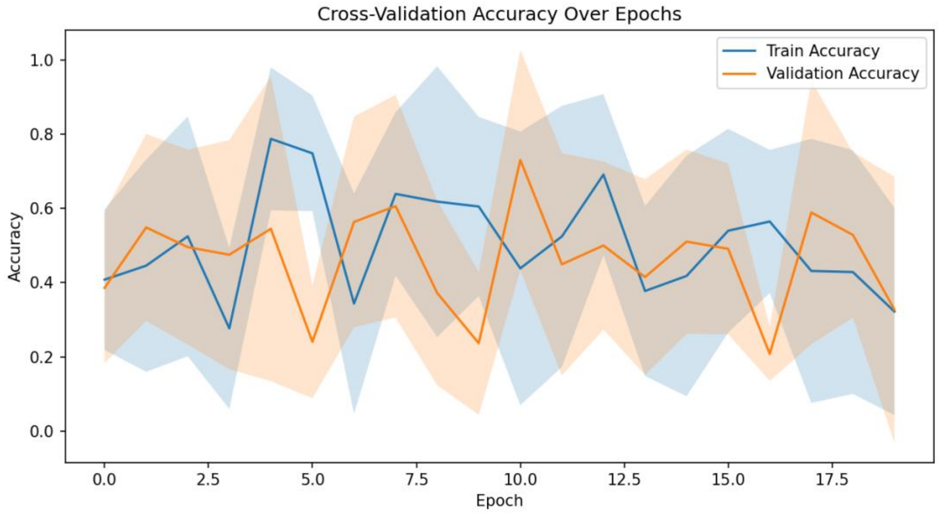


Figure 5. Cross-validation accuracy over epochs. The solid lines represent the average value while the shaded areas indicate variability within ±1 standard deviation among the folds.

The learning curves associated with stratified 5-fold cross-validation are depicted in Figure 5, showing the average training accuracy (blue) and validation accuracy (orange) across epochs. In this scheme, the dataset is divided into 5 equally sized folds, with the model trained on 4/5 of the data and validated on the remaining 1/5 for each fold. This approach allows for an unbiased assessment of model performance on unseen data, as the held-out portion is not utilized during training. Performance metrics, such as accuracy, are calculated on this validation set at each epoch, and the results are averaged across all folds to provide a robust estimate of model generalization.

The learning curves illustrate that the training accuracy fluctuates between approximately 0.2 and 0.8 over the course of training, rather than rising smoothly. The validation accuracy similarly varies from about 0.2 up to 0.7 across all epochs, reflecting significant fold-to-fold variability. Unlike typical curves, both lines show persistent fluctuations throughout the epochs rather than stabilizing for an extended period. The generalization gap—the difference between training and validation accuracy—changes unpredictably, at times exceeding 0.2, which indicates periods of both underfitting and possible overfitting depending on epoch and fold composition.

Shaded regions around the curves represent the variability in accuracy across different folds, highlighting substantial differences in performance that arise from the biological and sampling diversity present in the dataset. Since no separate test set was used, all accuracy metrics are based exclusively on the cross-validation results. This approach leads to an estimated model accuracy that fluctuates between 20% and 80%, indicating moderate yet highly variable predictive ability and inconsistent learning patterns across the dataset. The validation accuracy curve varies from 0.2 to 0.7, meaning the model correctly predicts the outcome for 20–70% of samples in unseen validation folds during cross-validation. In parallel, the training accuracy ranges from about 0.2 to 0.8, reflecting notable fluctuations rather than steady improvement over epochs. These metrics provide a reliable estimate of the expected model performance when applied to new, unseen data, aligning with standard practices in machine learning where cross-validation results are reported as averages to ensure robust error estimates.

From both biological and modeling perspectives, several significant observations emerge. The relatively modest generalization gap suggests that the model adeptly captures genomic patterns specific to hosts without merely memorizing the details of the training data. Additionally, the considerable inter-fold variability serves as a reflection of data limitations, particularly concerning minority classes, thereby highlighting the pressing need for improved representation or targeted data augmentation strategies.

The model's ability to achieve approximately 20-70% validation accuracy, even within a context marked by high imbalance and biological diversity, implies that meaningful host-specific k-mer signals exist, particularly considering that only nucleotide data for genotypes 3 and 4 was employed. While these results are encouraging, caution is warranted against overinterpretation. The model exhibits proficient generalization for dominant hosts, such as humans and pigs. However, its capacity to accurately classify rare or zoonotic spillover occurrences, such as those involving deer or wild boars, is constrained due to the limited availability of data, a limitation reflected in the broad variability observed across the folds.

To enhance the robustness of the model and improve its predictive capabilities, several recommendations can be made. First, it is advisable to continue utilizing the stratified cross-validation framework and provide mean ± SD accuracy at pivotal epochs to ensure transparency in performance reporting. Second, the integration of class-weighted loss functions or focal loss should be considered to better address the challenges posed by class imbalance, promoting a more equitable learning process.

Furthermore, options for minority-class augmentation or synthetic data generation should be explored to stabilize the performance of validation metrics, ensuring that the model remains resilient in the face of data scarcity.

4.5. Overall predictive performance

The performance of the final neural network model was rigorously evaluated through a detailed analysis of the confusion matrix and the recall metrics for each host class. These tools provide vital insights into how well the model classifies different categories of hosts associated with the Hepatitis E virus, thus shedding light on its predictive capabilities.

Figure 6 illustrates the confusion matrix, assessed on the test dataset, that is a critical component in assessing model performance. In this matrix, the true labels of host categories are represented along the vertical axis, while the predicted labels generated by the model are displayed along the horizontal axis. The intersections of these rows and columns reveal the frequency of correct and incorrect predictions.

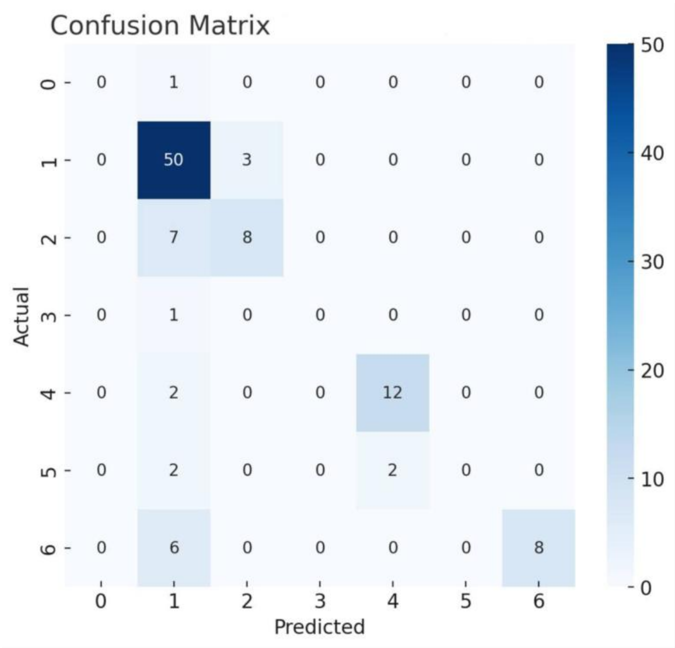


Figure 6. Confusion Matrix

The model demonstrates strong performance in predicting the human (encoded as 1) and pig (encoded as 4) hosts, achieving 50 and 12 correct predictions, respectively. These results are reflected in the prominent diagonal entries for these classes, indicating that the classifier effectively distinguishes hosts with distinct characteristics.

In addition to its strengths in predicting the human (encoded as 1) and pig (encoded as 4) hosts, the model shows moderate accuracy for wild boar (encoded as 6), with 8 out of 14 samples correctly classified as wild boar. However, 6 wild boar samples are misclassified as human, indicating the model may confuse wild boar features with those of human or that these classes share overlapping characteristics in the feature space.

For the other categories, model weaknesses become apparent. Environment (encoded as 2) is misclassified as human seven times and shows only eight correct predictions, while rabbit (encoded as 3) and deer (encoded as 0) are solely misclassified as human, with one sample each. This recurring pattern - multiple non-human hosts being classified as human - highlights the impact of possible class imbalance and overlapping feature distributions, which confound the model's accuracy for less represented categories.

In addition to the confusion matrix, Figure 7 presents the recall per class, which quantifies the sensitivity of the model in correctly identifying instances of each host category. Recall is defined as the ratio of true positives to the sum of true positives and false negatives, providing a clear measure of the model's performance for each specific class.

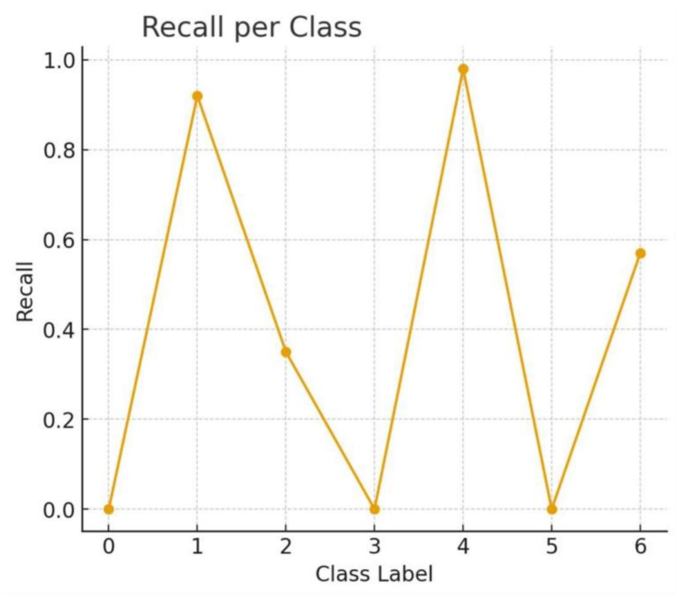


Figure 7. Recall per Class

Pig (class 4) achieves a perfect recall of 1.0, indicating that all sequences from this host were correctly identified. Human (class 1) also shows strong performance, with recall close to 0.9.

By contrast, several host categories, including deer (class 0) and rabbit (class 3), have a recall of 0.0, meaning the model failed to identify any instances from these classes. Environment (class 2) and wild boar (class 6) exhibit moderate recall values of around 0.37 and 0.57, respectively, suggesting partial but inconsistent recognition. These results highlight both the strengths of the model in recognizing certain hosts and its limitations in distinguishing others.

The combined insights from the confusion matrix and recall analysis paint a comprehensive picture of the model's performance in the context of zoonotic risk assessment for Hepatitis E virus. While the model excels in accurately identifying well-represented host classes, it struggles significantly with rare or closely related categories, which poses challenges for effective monitoring and prediction.

Understanding these limitations is crucial for future improvements. The identification of misclassification patterns suggests that further data collection efforts, particularly for the underrepresented classes, could enhance the model's predictive accuracy. Additionally, employing advanced techniques such as data augmentation or feature engineering may help mitigate the effects of overlapping features and improve the model's ability to distinguish between complex host categories.

In summary, the evaluation of the confusion matrix and recall metrics not only highlights the strengths of the model but also emphasizes areas for further investigation and refinement. The findings underscore the importance of continuous efforts in enhancing predictive models for zoonotic diseases, as they play a vital role in public health surveillance and prevention strategies.

4.5.1. Error analysis and what the model is learning

A notable strength of the model lies in its high recall rates for the dominant classes. This observation suggests that host-specific evolutionary pressures impart distinct compositional signatures within the viral genome. For HEV genotypes 3 and 4, which predominantly circulate in swine populations before spilling over into human hosts, these identifiable signals likely emerge due to various factors. Differences in the pools of host tRNA, the influence of innate immune responses such as biases introduced by APOBEC and ADAR enzymes, and the specific environments in which replication and translation occur play significant roles in shaping codon and k-mer usage patterns. Phenotypic changes related to receptor affinity or to host-specific immune peculiarities might also indirectly be reflected in the genome composition.

The neural network appears to capitalize on these stable genomic signatures effectively. By recognizing these evolutionary cues, the model demonstrates its ability to classify the dominant classes accurately, reflecting the underlying biological realities of HEV transmission dynamics.

Conversely, the model faces significant hurdles when dealing with classes represented by sparse sequences, such as those from rabbits, capybaras, sporadic wildlife, and environmental samples. The recall rates for these classes hover near zero, indicating a pronounced difficulty in accurate classification. Two primary mechanisms are suspected to contribute to this issue:

1. Statistical Insufficiency: The limited number of available examples for these classes hampers the model's ability to establish reliable decision boundaries within the 5-mer space. Without sufficient training data, the network struggles to learn the nuanced patterns necessary for effective classification.
2. Biological Overlap: In addition to accurately predicting human (encoded as 1) and pig (encoded as 4) hosts, the model often misclassifies environmental (encoded as 2) samples

as human. This tendency likely reflects a compositional overlap, potentially due to contamination from human waste in environmental sources. Such biological and environmental interactions can blur the boundaries between categories, complicating the classification process and highlighting the influence of cross-contamination on the model's decision-making.

From a public health perspective, these biased errors remain valuable. They provide critical information regarding the most likely source reservoirs for potential zoonotic transmission, even when the precise classification remains uncertain. This insight underscores the model's utility in guiding further investigations into HEV transmission pathways and the identification of potential risks associated with zoonotic diseases.

4.5.2. Robustness check and practical improvement

Improving data quality and representativeness is important. Increasing the number of sequences for under-represented hosts, such as wild boar, rabbits, and various non-suid wildlife species, should be a primary focus. In instances where immediate sequencing is not feasible, alternative methods like class - balanced sampling and the implementation of focal loss can effectively address data imbalance during the training phase. These techniques ensure fair representation of all host types.

Implementing hierarchical labeling can yield substantial benefits. A two-stage classification system could be adopted, where the initial step distinguishes between suid and non-suid species, followed by a more detailed classification into subclasses such as human, pig, wild boar, and other wildlife. This approach aligns with biological hierarchies and has been shown to enhance the performance of minority classes, resulting in improved overall accuracy.

Incorporating richer sequence features could greatly enhance model performance. Combining 5-mers with codon-usage indices, dinucleotide odds ratios, or utilizing one-hot encoding in convolutional neural networks trained on sliding windows can unveil structural details that simpler composition-based methods might overlook. A multi-faceted approach will bolster the model's ability to detect complex biological patterns.

Lastly, the concept of genotype-aware multitask learning should be considered. By predicting differences in genotype - specifically between types 3 and 4 - as an auxiliary task, the model can be better regulated. This strategy not only captures intricate interactions between host characteristics and genotype variations but also strengthens the overall predictive capabilities of the model.

In conclusion, while the initial findings of this study are indeed promising, numerous opportunities for improvement and refinement exist. By adopting these strategies, the robustness and accuracy of the zoonotic risk assessment for Hepatitis E Virus can be significantly enhanced.

5. CONCLUSION

This thesis has effectively examined the role of neural networks in identifying the origin of the hepatitis E virus through a thorough bioinformatic approach. By leveraging a carefully curated dataset of viral genomes alongside host metadata, the research highlights the promise of computational methods in overcoming challenges associated with host identification, especially in scenarios where genomic databases may be incomplete or unclear.

The development of a single-layer neural network model has yielded encouraging results in classifying viral sequences, demonstrating the power of deep learning techniques to handle extensive biological datasets. A notable finding of this study is that the method's ability to predict hosts based on genetic data suggests that host tropism is at least partially influenced by genetic factors. This insight underscores the need for further investigations, including experimental studies, to thoroughly characterize these determinants.

While the findings are promising, they also open avenues for future enhancements in virology, particularly in refining public health strategies to address hepatitis E outbreaks. This research contributes to a more nuanced understanding of virus transmission dynamics, highlighting the crucial role of reliable data and advanced modeling techniques in combating foodborne viral infections.

Despite significant advancements in understanding hepatitis E over recent decades, several key challenges persist in current research. These include the virus's genetic diversity and its transmission pathways. Additionally, the limited number of documented hepatitis E cases complicates efforts in this field. Most available data pertains to cases in Asia and Europe; therefore, acquiring more comprehensive data from North and South America, Africa, and Eastern Europe is essential to gain a complete understanding of the situation and the zoonotic potential of the virus.

6. BIBLIOGRAPHY

Alcami, A., & Koszinowski, U. H. (2000). Viral mechanisms of immune evasion. *Immunology today*, 21(9), 447–455. https://doi.org/10.1016/s0167-5699(00)01699-6

Baha, S., Behloul, N., Liu, Z., Wei, W., Shi, R., & Meng, J. (2019). Comprehensive analysis of genetic and evolutionary features of the hepatitis E virus. *BMC genomics*, 20(1), 790. https://doi.org/10.1186/s12864-019-6100-8

Caliskan, A., Dandekar, T., and Dangwal, S. "Big Data and Metadata in Scientific Research: Challenges and Opportunities." *Computational and Structural Biotechnology Journal*, vol. 21, 2023, pp. 4895-4913, https://doi.org/10.1016/j.csbj.2023.10.006

Ejaz, M., Zohaib, A., Usman, M., Anwar, M. M., Khan, H. S., Ammar, M., Saqib, M., Khan, S. G., Athar, M. A., Mansoor, M. K., Ullah, A., Naseem, M., Hussain, M. H., & Javed, A. (2024). Genotypic analysis of hepatitis E virus (HEV) from sporadic symptomatic cases in Pakistan. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 118, 105567. https://doi.org/10.1016/j.meegid.2024.105567

Elois, M. A., Pavi, C. P., Jempierre, Y. F. S. H., Pilati, G. V. T., Zanchetta, L., Grisard, H. B. D. S., García, N., Rodríguez-Lázaro, D., & Fongaro, G. (2025). Trends and Challenges in the Detection and Environmental Surveillance of the Hepatitis E Virus. *Microorganisms*, 13(5), 998. https://doi.org/10.3390/microorganisms13050998

Erickson, B. J., and Kitamura, F. "Magician's Corner: 9. Performance Metrics for Machine Learning Models." Radiology: Artificial Intelligence, vol. 3, no. 3, 2021, e200126, https://doi.org/10.1148/ryai.2021200126

Forzan, M., Pacini, M. I., Pericciolі, M., & Mazzei, M. (2021). Hepatitis E Virus RNA Presence in Wild Boar Carcasses at Slaughterhouses in Italy. *Animals : an open access journal from MDPI*, 11(6), 1624. https://doi.org/10.3390/ani11061624

Garbuglia, A. R., Bruni, R., Villano, U., Vairo, F., Lapa, D., Madonna, E., Picchi, G., Binda, B., Mariani, R., De Paulis, F., D'Amato, S., Grimaldi, A., Scognamiglio, P., Capobianchi, M. R., Ciccaglione, A. R., & The Other Members Of The Hev Outbreak Working Group (2021). Hepatitis E Outbreak in the Central Part of Italy Sustained by Multiple HEV Genotype 3 Strains, June-December 2019. *Viruses*, 13(6), 1159. https://doi.org/10.3390/v13061159

Garbuglia, A. R., Bruni, R., Villano, U., Vairo, F., Lapa, D., Madonna, E., Picchi, G., Binda, B., Mariani, R., De Paulis, F., D'Amato, S., Grimaldi, A., Scognamiglio, P., Capobianchi, M. R., Ciccaglione, A. R., & The Other Members Of The Hev Outbreak Working Group (2021). Hepatitis E Outbreak in the Central Part of Italy Sustained by Multiple HEV Genotype 3 Strains, June-December 2019. *Viruses*, 13(6), 1159. https://doi.org/10.3390/v13061159

Iaconelli, M., Bonanno Ferraro, G., Mancini, P., Suffredini, E., Veneri, C., Ciccaglione, A. R., Bruni, R., Della Libera, S., Bignami, F., Brambilla, M., De Medici, D., Brandtner, D., Schembri, P., D'Amato, S., & La Rosa, G. (2020). Nine-Year Nationwide Environmental Surveillance of Hepatitis E Virus in Urban Wastewaters in Italy (2011-2019). *International journal of environmental research and public health*, 17(6), 2059. https://doi.org/10.3390/ijerph17062059

Kamani, L., Padhani, Z. A., & Das, J. K. (2021). Hepatitis E: Genotypes, strategies to prevent and manage, and the existing knowledge gaps. *JGH open : an open access journal of gastroenterology and hepatology,* 5(10), 1127–1134. https://doi.org/10.1002/jgh3.12646

Khuroo, M. S., Khuroo, M. S., and Khuroo, N. S. "Transmission of Hepatitis E Virus in Developing Countries." *Viruses*, vol. 8, no. 9, 2016, art. 253, https://doi.org/10.3390/v8090253

Matuszewski, D. J., and Sintorn, I.-M. "TEM Virus Images: Benchmark Dataset and Deep Learning Classification." Computer Methods and Programs in Biomedicine, vol. 209, 2021, 106318, https://doi.org/10.1016/j.cmpb.2021.106318

Mirazo, S., and Arbiza, J. "Hepatitis E and Chronic Liver Damage in Apparently Immunocompetent Individuals: Now What?" *Annals of Hepatology*, vol. 18, no. 3, 2019, pp. 539–540, https://doi.org/10.1016/j.aohep.2019.05.002

Moeckel, C., Mareboina, M., Konnaris, M. A., Chan, C. S. Y., Mouratidis, I., Montgomery, A., Chantzi, N., Pavlopoulos, G. A., and Georgakopoulos-Soares, I. "A Survey of K-mer Methods and Applications in Bioinformatics." Computational and Structural Biotechnology Journal, vol. 23, 2024, pp. 2289–2303, https://doi.org/10.1016/j.csbj.2024.05.025

Purdy, M. A., & Khudyakov, Y. E. (2010). Evolutionary history and population dynamics of hepatitis E virus. *PloS one,* 5(12), e14376. https://doi.org/10.1371/journal.pone.0014376

Reiman, D., Metwally, A. A., Sun, J., and Dai, Y. "PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype from Metagenomic Data." IEEE Journal of Biomedical and Health Informatics, 2020, https://doi.org/10.1109/JBHI.2020.2993761

Rein, D.B. et al. (2021). Hepatitis E virus: advances and challenges. *Trends in Microbiology*, 29(4), 312-319. https://doi.org/10.1016/j.tim.2021.01.004

Ren, J., Bai, X., Lu, Y. Y., Tang, K., Wang, Y., Reinert, G., and Sun, F. "Alignment-Free Sequence Analysis and Applications." Annual Review of Biomedical Data Science, vol. 1, 2018, pp. 93–114, https://doi.org/10.1146/annurev-biodatasci-080917-013431

Songtanin, B., Molehin, A. J., Brittan, K., Manatsathit, W., and Nugent, K. "Hepatitis E Virus Infections: Epidemiology, Genetic Diversity, and Clinical Considerations." *Viruses*, vol. 15, no. 6, 2023, 1389, https://doi.org/10.3390/v15061389

Tong, H., Hoan, N. X., Wang, B., Wedemeyer, H., Bock, C. T., & Velavan, T. P. (2016). Hepatitis E Virus Mutations: Functional and Clinical Relevance. *EBioMedicine*, 11, 31–42. https://doi.org/10.1016/j.ebiom.2016.07.039

Wang, L., Teng, J. L. L., Lau, S. K. P., Sridhar, S., Fu, H., Gong, W., Li, M., Xu, Q., He, Y., Zhuang, H., Woo, P. C. Y., and Wang, L. "Transmission of a Novel Genotype of Hepatitis E Virus from Bactrian Camels to Cynomolgus Macaques." *Journal of Virology*, vol. 93, no. 7, 2019, e02014-18, https://doi.org/10.1128/JVI.02014-18

Wang, B., and Meng, X.-J. "Hepatitis E Virus: Host Tropism and Zoonotic Infection." Current Opinion in Microbiology, vol. 59, 2021, pp. 8–15, https://doi.org/10.1016/j.mib.2020.07.004

Yang, L. "Theoretical Analysis of Adam Optimizer in the Presence of Gradient Skewness." International Journal of Applied Science, vol. 7, no. 2, 2024, pp. 27–42, https://doi.org/10.30560/ijas.v7n2p27